MODEL SENSITIVITY TO PRIOR SELECTION IN REPLICATION STUDIES

BY

QIAWEN LIU

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Educational Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Adviser:

Professor Carolyn Anderson

ABSTRACT

When doing a Bayesian Analysis for a replication study, selecting priors is a widely discussed issue. On one hand, we could argue that an informative prior specified by previous research is preferable because we have some knowledge and expectations regarding the phenomena. However, when the goal is to replicate findings from previous research, we do not want to use prior findings to influence results of the replication study; that is, for a replication study, we should use a non-informative prior, which would maximize the utility of current data. By analyzing a replication research for a widely cited psycholinguistics paper (Fine, Jaeger, Qian, & Farmer, 2013), this thesis aims to provide insight as to how a replication researcher might go about selecting priors for analyzing replication studies within a Bayesian framework. By using sensitivity analyses, posterior predictive checking, and information criteria, researchers can start with a more reasonable prior setting that eventually leads to more valid confirmation or non-confirmation of previous research.

ACKNOWLEDGEMENTS

I would like to express the greatest appreciation to my committee chair, Professor Carolyn Anderson, who wisely and patiently guide me throughout the writing of this thesis. Without her guidance and help this dissertation would not have been possible.

I would like to thank my committee members, Professor Kiel Christianson and Professor Jinming Zhang. The thesis is an integral part of a psycholinguistics project that I have been working on with Professor Christianson from 2018 to 2019. He leads me into the wonderland of psycholinguistics, to which I would devote my next five years of PhD study and even the rest of my academic life. Meanwhile, I would like to thank Professor Jinming Zhang for providing his valuable feedback and suggestions to this thesis.

My sincere thanks also go to my co-worker and friends Jack, Marian, Wenying, and Yinhao. They not only kindly contribute to this project, but also keep me company on long walks. Their encouragement and camaraderie endowed me with magnificent power to overcome every obstacle in pursuing this Master's degree.

Last, and most of all, I would like to thank my parents for providing me with unfailing support throughout my years of study. This thesis would not have been possible without their endless love.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

In the field of psychology, small sample sizes and underpowered studies are endemic. Combined with the publication bias which overstates the statistical significance threshold (p <.05) and only the studies with large significant effect are accepted for publication. The field is rife with false alarms and suffers an ever-lasting replication crisis. Judging from a traditional Null-Hypothesis Significance Testing (NHST) perspective, according to the Reproducibility Project: Psychology by the Open Science Collaboration (OSC), only 39% of nearly 100 studies reached statistical significance and were said to have successfully replicated the original study. However, it would be too hasty to conclude that all these "failures" to replicate indicate that the original results are wrong, since it is also acknowledged that the .51 correlation of effect sizes between the replication and the original show a moderate robustness of the original results.

Though NHST remains to be the field's most widely used method when it comes to evaluating hypotheses, it may fall short in evaluating replication studies. In NHST, the analysis usually starts with a set of very constrained computational assumptions and a point estimate of parameters (no other plausible parameter values). More importantly, researchers either reject or fail to reject the null hypothesis in frequentist analyses, and such binary decision-making provides little information about the quantified evidence in favor of the null hypothesis. Therefore, if a statistical test fails to reject the null hypothesis, researchers are left with little more to say. Obviously, the replication success can hardly be evaluated sufficiently through this method.

Meanwhile, with the recent advances in computational capacity and the availability of Bayesian estimation in widely-used computer software, probabilistic programs such as Stan (Stan Development Team, 2018), and accessible packages in R environment (R Core Team,

2014) like brms (Bürkner, P. C., 2016), there has been a steady increase in the application of Bayesian statistical methods across all fields of scientific research. Replication studies could benefit greatly from using Bayesian methods. Compared to NHST, the Bayesian model could easily adapt to the specific circumstance without as many computational restrictions, which are pervasive in NHST approaches. In addition, some complex models (e.g., mixture, multilevel, or longitudinal modeling) seem to benefit from Bayesian methods when it comes to convergence issues (Depaoli, & Van de Schoot, 2017; van Loey, & Sijbrandij, 2015; Depaoli & Clifton, 2015; Rabe-Hesketh & Skrondal, 2012), generating more accurate parameter estimations (Depali, 2013), and model specifications (Kim, Suh, Kim, Albanese, & Langer, 2013), and such models have become more and more popular in the field of psychology for analyzing repeated measures or other nested data. Most importantly, these methods provide clear quantified results as to the extent to which the data support either hypothesis by directly evaluating the strength of evidence for the null and alternative hypotheses.

However, applying Bayesian methods can be challenging for several reasons. First and foremost, Bayesian estimation requires making use of background information (subjective priors), while posterior distribution could be considered as a weighted average of likelihood from data and the prior distribution.

$$\text{Posterior} \propto \text{Likelihood} \cdot \text{Prior}$$

The prior distribution models the uncertainty and relative credibility of the parameter values before new data are considered. Once new data enter the equation, the posterior distribution becomes a compromise between the prior distribution and the likelihood of the parameter value suggested by the data. Incorporating reasonable and informative priors is ideal as they make evidence cumulative and reflect steady scientific progress. However, if

inappropriate informative priors are chosen, the posterior distribution can be influenced dramatically by this mis-specification. Meanwhile, if the prior distribution is vague and weakly informative, the posterior distribution is usually steered by the data and is less affected by the prior. There have therefore been debates about whether replication studies should use informative or noninformative/weakly informative priors.

The current study aims to evaluate Bayesian models using different priors of an experiment seeking to replicate a widely cited psycholinguistics paper (Fine, Jaeger, Qian, & Farmer, 2013). The results from their original experiments ostensibly showed that readers rapidly adapt their expectations to match novel syntactic frequency distributions. However, Harrington-Stack, James, & Watson (2018) conducted a 95% power replication study and failed to replicate this finding. Given this uncertain adaptation effect, it's a remained question whether we should use an informative prior specified by the original theory.

CHAPTER 2: LITERATURE REVIEW

2.1 MODEL SENSITIVITY

Prior distribution usually falls into three main categories, depending on how uncertain the researcher is about that parameter value: 1) noninformative prior, 2) weakly informative prior, or 3) informative prior. Typically, a noninformative prior could be described by a relatively flat distribution under which the parameter values have approximately equal likelihood. (e.g., the prior Uniform (-10, 10) is a straight line parallel to the x-axis). If a prior distribution contains some useful information but would not affect the resulting posterior estimate too much, the priors could be referred to as weakly informative. In some ways, a weakly informative prior is more functional than noninformative priors because inappropriate inferences drawn from a noninformative prior could be avoided (Gelman, Jakulin, Pittau, & Su, 2008). For example, if you know the parameter is likely to have a negligible effect, you can specify a weak informative prior Normal(0, 100) where 0 is the mean and 100 is the standard deviation, as the distribution is so spread out that the parameter value is completely ambiguous. A generic weakly informative prior is more informative than the above ones by specifying a prior with smaller variance but meanwhile large enough to cover the possible true values. For example, if you know that a negative starting point is unlikely for your model, you would probably specify a positive value for your intercept but still allow for a relatively wide fluctuation around the mean (e.g., Normal (10, 5). An Informative prior distribution could alternatively be specified, containing definite numerical values that reflect to a high degree of certainty estimates for the model's distribution. Vanpaemel (2010) advocates for using informative priors because they are vehicles for expressing psychological theory and should be considered as an integral part of the model to advance knowledge cumulatively. There are several strategies to elicit an informative prior for a

replication study (Depaoli & Schoot, 2017). First, an expert in the field could give an empirical estimate of the hyper-parameters (Bijak & Wisniowski, 2010; Fransman et al., 2011; Howard, Maxwell, & Fleming, 2000; Martin et al., 2012; Morris, Oakley, & Crowe, 2014). Second, the researcher conducting the replication could utilize the results of the original study as priors (Kaplan & Depaoli, 2013). Third, by combining multiple studies, the researcher could conduct a meta-analysis and use the results to define hyper-parameter values for the prior (Ibrahim, Chen, & Sinha, 2005; Rietbergen et al., 2011). For example, Ostarek et al. (2018) applied informative prior estimates based on previous experiments using the sentence-picture verification paradigm (Rommers, Meyer, & Huettig, 2018; Zwaan & Pecher, 2012; Zwaan et al., 2002) while doing a Bayesian follow-up analysis to investigate the amount of evidence in favor of the target effect. Fourth, the researcher could conduct a pilot study with a similar population of interest and implement a sampling method to estimate the parameter that could be used to define the priors for the subsequent studies (Gelman, Bois & Jiang, 1996). Finally, the researchers could use the maximum likelihood (cf. Berger, 2006; Brown, 2008; Candel & Winkens, 2003; van der Linden, 2008) or sample statistics (e.g., Darnieder, 2011; Raftery, 1996; Richardson & Green, 1997; Wasserman, 2000) to derive priors; however, such methods are criticized for "double-dipping" (Darnieder, 2011) because sample data are first used for generating priors and then used again for estimation.

Priors can substantively influence research findings, especially when sample sizes are small (Depaoli, 2013; Gelman & Shalizi, 2013; Johnson, 2013; Seaman, Seaman, & Stamey, 2012; van de Schoot & Depaoli, 2014). The critical question is to what extent these subjective priors influence the output from Bayesian analyses, and whether such influence is warranted. It stands to reason that specifying a correct informative prior has an ideal impact on the posterior

distribution while specifying an inappropriate informative prior can skew the results. However, perhaps less intuitively, a non-informative prior could also act as an informative prior (Gelman, 2006a). For example, specifying a Dirichlet prior (10, 10) for a two-class mixture model can shape the posterior and push two classes to be equal even if in reality they are far apart. Also, a non-informative prior might actually become unintendedly informative if the parameter is transformed (e.g., logit transformation in logistic regression) (Seaman et al. 2012).

Since priors might have a large impact on posterior estimates, and the effect of the priors are usually uncovered using various diagnostic tools, it is important to understand prior specification and evaluate different priors carefully before interpreting results from the Bayesian analysis. The following section will introduce several ways to evaluate models using different priors.

2.2 EVALUATING METHODS

2.2.1 Posterior predictive checking

Gelman, Meng, and Stern (1996) proposed a Bayesian approach for conducting a goodness of fit assessment, namely posterior predictive checking, which directly measures the discrepancy between the observed data and the fitted model through the posterior simulation. Using y for the observed data, θ for parameters, and y^{rep} for the simulated data from the predictive distribution, the relationship is expressed as:

$$p(y^{rep}|y) = \int p(y^{rep}|\theta, y)p(\theta|y)d\theta.$$

First, we simulate m values ofθ from the posterior distribution, $p(\theta|y)$. Then, each y^{rep} is simulated from the likelihood $p(y^{rep}|\theta,y)$. Then, we compare y to the replicated datasets y^{rep}. Visual discrepancy tests could be conducted to examine the residuals of their expectations under a fitted model.

Using posterior predictive checking, we can compare the posterior predictive distribution of models under different prior settings to the observed data. If we observe some models that particularly deviate from the data in this step, that could be an indication of an inappropriate prior setting.

2.2.2 Sensitive analysis

A sensitivity analysis of priors is recommended when informative or weakly informative priors are applied, as there might be a discrepancy between estimations using different subjectively chosen priors. A sensitivity analysis usually involves re-estimating the model after adjusting the entire prior distribution or increasing/decreasing hyperparameters of a certain prior setting. The scale of adjustment is specific to the data. For example, if the initial mean is specified at 31.37, Depaoli and Schoot (2017) suggested that researchers examine a series of priors with mean hyperparameters in 5-point increments/decrements from the initial mean (i.e., 21.37, 26.36, 31.37, 36.37, 41.37). Then, using the Gelman and Rubin convergence diagnostic, i.e., Potential scale reduction factor (PSRF), which estimates the potential decrease in the between-chains variability with respect to the within-chain variability, researchers examine whether the chains from these priors substantially deviate from each other.

Also, the difference between size of the effects can be computed to assess the extent to which the results of models with different priors means (Depaoli & Schoot, 2017). The relative deviation of size of the effect is computed as:

(estimate using subjective prior - estimate using new prior) / (estimate using subjective prior) × 100%

If the difference (e.g., percent of relative deviation) is low enough (e.g., under 1% for relative deviation), then the results could be considered relatively stable with the use of different mean

hyperparameters. Researchers can continue this sensitivity analysis with the variance hyperparameter to see how the diagnostic statistics vary across prior selections.

If even a small fluctuation in hyperparameter values would cause great instability in substantive results, this could be an indication of model mis-specification or some parameters are mis-identified.

2.2.3 Information criterion

Leave-one-out cross-validation (LOO) and Widely Applicable Information Criterion (WAIC) are methods for estimating point-wise prediction accuracy from a fitted Bayesian model. They have been widely used for the purpose of model comparison, selection, or averaging (Ando & Tsay 2010; Geisser & Eddy 1979; Hoeting et al. 1999; Vehtari & Lampinen 2002; Vehtari & Ojanen 2012). Both of these methods use the log-likelihood from the posterior simulations of the parameters to estimate out-of-sample predictive accuracy.

Consider independent data set of size n: $y_1, \ldots, y_n$, and suppose we have a prior distribution $p(\theta)$, thus yielding a posterior distribution $p(\theta|y)$ and a posterior predictive distribution $\int p(\tilde{y}_i|\theta)p(\theta|y)d\theta$.

elpd (expected log pointwise predictive density)

$$= \sum_{i=1}^{n} \int p_t(\tilde{y}_i) \log p(\tilde{y}_i|y)\, d\tilde{y}_i \,,$$

where $p_t(\tilde{y}_i)$ is the distribution representing the true data-generating process for $\tilde{y}_i$. The $p_t(\tilde{y}_i)'s$ are unknown, and we will use cross-validation or WAIC to approximate elpd.

The LOO cross validation estimate of out-of-sample predictive density is the summation of log predictive density given the data without the ith point

$$elpd_{loo} = \sum_{i=1}^{n} \log p(y_i|y_{-i}) \ ,$$

where the conditional probability $p(y_i|y_{-i})$ is defined as

$$p(y_i|y_{-i}) = \int p(y_i|\theta)p(\theta|y_{-i})d\theta.$$

Vehtari, Gelman, and Gabry (2016) proposed *Pareto smoothed importance sampling (*PSIS) to give a stable estimate of LOO, and the brms package in R provides a handy estimate of LOO based on this method.

WAIC could be considered as an improvement on the deviance information criterion (DIC), which is also an alternative method of estimating the expected log pointwise predictive density (elpd). Though DIC has been gaining popularity recently, it is known to have problems in evaluating Bayesian models because it is based on a traditional point estimate (van der Linde 2005; Plummer 2008). For instance, DIC is undefined for singular models where the covariance matrices are singular and can possibly produce negative estimates for an effective number of parameters. In contrast, WAIC is generated from the entire posterior distribution and asymptotically equivalent to Bayesian cross-validation. Also, it is invariant to parameterization and works for singular models (Vehtari, Gelman & Gabry, 2016).

$$\widehat{elpd}_{waic} = \widehat{lpd} - \hat{p}_{waic},$$

where

$$\widehat{lpd} = \text{computed log pointwise predictive density} = \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\theta^s)\right),$$

and

$$p_{waic} = \sum_{i=1}^{n} var_{post}(\log p(y_i|\theta)).$$

Both LOO and WAIC are methods to estimate prediction error, therefore, we can select the “best” model by minimizing LOO or WAIC.

CHAPTER 3: CASE STUDY

3.1 ORIGINAL STUDY & CURRENT REPLICATION

In psycholinguistics research, so-called garden-path (GP) sentences (Bever, 1970) are a broadly investigated phenomenon. People who encounter garden-path sentences are tricked into generating a syntactic representation of the sentence based on early comprehension that must later be corrected after additional information is encountered that does not fit with the initial syntactic parse. For example, in *The experienced soldiers warned about the dangers conducted the midnight raid*, readers initially read the sentence as the soldiers warning about the dangers in the past tense until they come across the disambiguating verb *conducted* and must reparse the global syntax to generate a correct representation: the soldiers who **were warned** about the dangers conducted the midnight raid (Frazier & Rayner, 1982). In the real world, such reduced-relative clauses (RC) GP sentences are much less common then their main verb (MV) continuation counterparts (e.g., the experienced soldiers warned about the dangers before the midnight raid), and readers will expect the verb *warned* to be a main verb due to this discrepancy in frequency (REFERENCE; Trueswell, Tanenhaus, & Garnsey, 1992, perhaps?). This ambiguity effect is usually reflected by some syntactic repair cost for reading RC sentences (e.g., slower reading times at the disambiguating word). Utilizing this long-observed pattern of processing behaviors, Fine, Jaeger, Farmer, and Qian (2013) conducted widely-cited study ostensibly showing that adult readers are able to rapidly adjust their syntactic expectations if they are exposed to statistically unbalanced input in the form of relative clause GP structures during the course of an experiment.

The current study sought to replication Fine et al.'s results using nearly identical materials and an identical design. Three research questions addressed in this paper (Fine et al., 2013) were

the focus of the current study. First, will participants who experience boosted exposure to RC sentences show lessened ambiguity effects reading ambiguous RCs during the course of an experiment (e.g., from block 1 to block 2)? Second, do these participants experience more difficulty processing ambiguous MV sentences after extensive exposure to RCs as a tradeoff compared to those who did not receive this exposure? Third, do early exposure group participants experience weaker ambiguity effects than those of the late group in the second block?

Two groups of participants (N=80) were recruited via Amazon Mechanical Turk. There are four possible types of sentences encountered in the experiment, depending on which group and block they are in (1-2).

1a. The experienced soldiers warned about the dangers conducted the midnight raid. (ambiguous RC)

1b. The experienced soldiers who were warned about the dangers conducted the midnight raid (unambiguous RC)

2a. The experienced soldiers warned about the dangers before the midnight raid. (ambiguous MV)

2b. The experienced soldiers spoke about the dangers before the midnight raid. (unambiguous MV)

Sentence (1a) and (2a) are temporarily ambiguous during the critical region (...*warned about the…)* but can be disambiguated when *conducted* is encountered. Sentence (1b) is unambiguous because *who* serves as a disambiguating cue, while sentence (2b) is also unambiguous because *spoke* could only be taken as a past tense intransitive matrix verb. The design of the experiment is summarized in Table 1. The Early (exposure) Group was exposed to RCs from the first block,

while the late (exposure) Group was not exposed to RCs until the second block. In the third block, both of the groups encountered 5 ambiguous MV (2a) and 5 unambiguous MV (2b). The hypotheses are: 1) The Early Group readers show faster reading times on ambiguous RC conditions in the second block comparing to the first block, i.e., there is an interaction between ambiguity and block. 2) In Block 3, the Early Group will be slower reading ambiguous MV sentences comparing to unambiguous MV sentences since they have adapted to their RC counterparts in the earlier two blocks, while the Late Group will not show such a difference between ambiguous and unambiguous MV, i.e., there is an interaction between group and ambiguity.

Table 1: Experimental Design by Fine et al. (2013)

	Block 1 (early group exposure starts)	**Block 2 (late group exposure starts)**	**Block 3 (RCs substituted by MVs)**
Early Group (#participant =40)	16 RCs (8 ambiguous)	10 RCs (5 ambiguous), 20 Fillers	10 MVs (5 ambiguous), 15 Fillers
Late Group (#participant =40)	16 Fillers	10 RCs (5 ambiguous), 20 Fillers	10 MVs (5 ambiguous), 15 Fillers

Fine et al. (2013) reported that early group participants significantly adapted to RCs. The current replication study slightly adapted materials from the original study (ensuring grammaticality and a lack of lexical ambiguity for all items), and added a comprehension

question after each RC and MV sentence directly probing the agenthood of the initial verb; in addition, 16 fillers were added to the first block for the Early Group in case the "adaptation effect" is in fact just a practice effect due to repeated rehearsal of the structure. No significant result was found for any parameter of interest in both questions in Fine et al.'s study. Consequently, Bayesian-follow up analyses were conducted to investigate the magnitude of the non-significant result. Since Question 2 and Question 3 are asking two similar questions, we are only discussing Question 1 and Question 2 in this thesis for the sake of concision. Bayesian models with different priors are evaluated and compared in the following sections.

3.2 CHOICE OF PRIOR

Four prior settings are used in the current study. Based on previous research, the average log reading time in msec on each word for self-paced reading is around 5.8, therefore, the intercepts for all models (except the one defined by Fine et al.'s results) are set around 5.8. For the parameters of interest (Question 1: interaction between Group and Ambiguity condition, Question 2: interaction between Block and Ambiguity), the first prior setting is very specific and informative, provided by the posterior distribution of the parameter in the original study (Fine et al., 2013). The second prior setting includes noninformative flat priors (uniform distribution over (-10, 10) for the parameter of interest). The third prior setting is weakly informative: normal ($\mu = 0$, $\sigma = 100$) prior, since the large standard deviation leads to a highly vague prior distribution. The last prior setting follows a Normal (0, 1), which is a generic weakly informative prior. The prior settings for two questions are summarized in the following tables.

Table 2: Prior summary for Question 1

Q1 prior summary	**noninformative**	**weakly informative**	**generic weakly informative**	**informative**
Intercept	Uniform (0,11.6)	Normal (5.8,100)	Normal (5.8,1)	Normal (5.7,0.03)
Ambiguity	Uniform (-10,10)	Normal (0,100)	Normal (0,1)	Normal (0.02, 0.01)
Group	Uniform (-10,10)	Normal (0,100)	Normal (0,1)	Normal (-0.02, 0.03)
Ambiguity x Group	Uniform (-10,10)	Normal (0,100)	Normal (0,1)	Normal (0.01, 0.01)

Table 3: Prior summary for Question 2

Q2 prior summary	**noninformative**	**weakly informative**	**generic weakly informative**	**informative**
Intercept	Uniform (0,11.6)	Normal (5.8,100)	Normal (5.8,1)	Normal (5.8,0.05)
Ambiguity	Uniform (-10,10)	Normal (0,100)	Normal (0,1)	Normal (0.03, 0.01)
Block	Uniform (-10,10)	Normal (0,100)	Normal (0,1)	Normal (-0.15, 0.03)
Ambiguity x Block	Uniform (-10,10)	Normal (0,100)	Normal (0,1)	Normal (-0.01, 0.01)

3.3 ANALYSIS

The Bayesian multilevel models were fit using the brms package in R (Bürkner, 2016), which uses the probabilistic programming language Stan. Stan implements Hamiltonian Monte Carlo (Duane, Kennedy, Pendleton, & Roweth 1987; Neal 2011) and its extension, the No-U-Turn Sampler (NUTS) (Hoffman and Gelman 2014). The alternatives, Gibbs sampling or Metropolis-Hastings updates, converge rather slowly for high-dimensional models with correlated parameters and require conjugate priors, Hamiltonian converges much more quickly (i.e., with fewer iterations) and does not require conjugate priors. Also, brms supports a wide

range of distributions and link functions, allowing users to fit multilevel models that could be easily converted to an mcmc object which is required form of input to many graphing and diagnostic purposes for Bayesian analysis.

Because reading times are skewed for our data, we log-transformed them so that they are more aligned with the assumption of a normal distribution (See Figure 1).

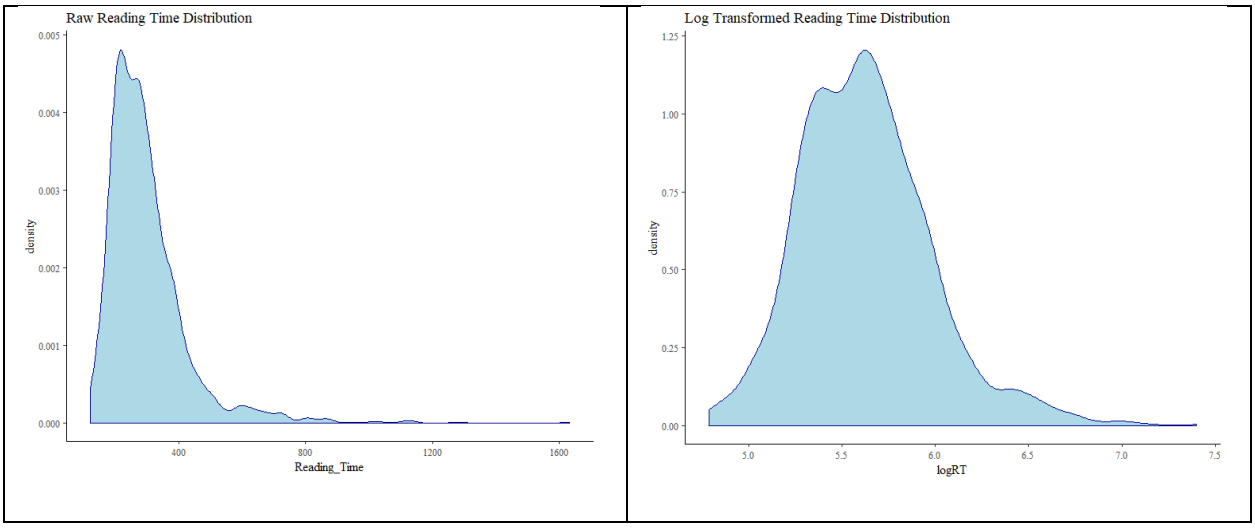


Figure 1: Raw reading time distribution (left) and log transformed reading time distribution (right)

The multi-level model for the first question is:

$$\log(Reading\ time_k) = Group_i + Ambiguity_i + Group_i \times Ambiguity_i + Word\ Length_k + U_{0j} + U_{1j} \times Ambiguity_i + W_{0k}.$$

The multi-level model for the second question is:

$$\log(Reading\ time_k) = Block_j + Ambiguity_i + Block_j \times Ambiguity_i + Word\ Length_k + U_{0j} + U_{1j} \times Ambiguity_i + U_{2j} \times Block_j + W_{0k} + W_{1k} \times Ambiguity_i,$$

where i is the index for group, j for participant, and k for item. U_{0j}, U_{1j} and U_{2j} are random intercept, random slope for Ambiguity and random slope for Block, grouped by Participant. W_{0k}

and W_{1k} are random intercept and random slope are random slope for Ambiguity, grouped by Item.

All brmfit objects were run for 4 chains, within each 1,000,000 iterations were run, the warm-up period is set to be the one-half number of the total iterations in each chain.

Before the result were analyzed, convergence was checked by examining autocorrelations and trace plots. As displayed in Figure 2 and Figure 3 which shows the autocorrelation and trace plot for the interaction term, the later lag autocorrelation becomes much smaller (i.e., close to zero) than the beginning point, and stably fluctuates around zero, indicating convergence. The trace plot shows the sampled values of a parameter over time and helps to judge how quickly the Markov Chain-Monte Carlo procedure converges in distribution. Also, in the figures, all the trace plots show rapid up-and-down variation with no apparent long-term trends or drifts, indicating convergence of all four chains. The convergence information for the rest of the parameters are included in the Appendix.

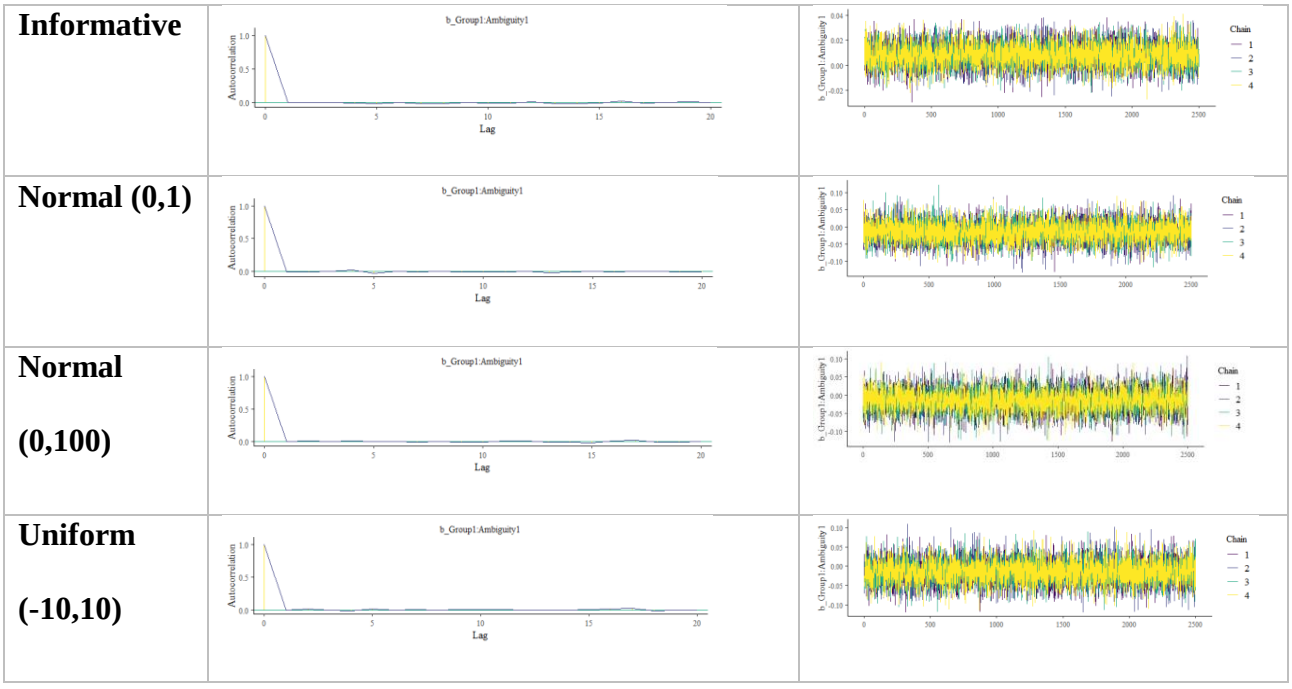


Figure 2: Autocorrelation and trace plot for Question 1, Group and Ambiguity interaction

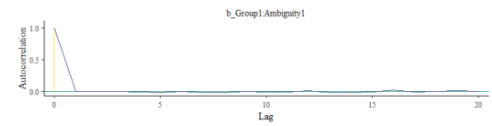
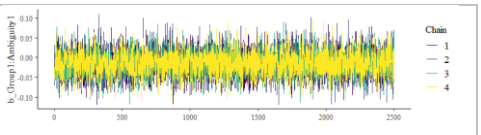
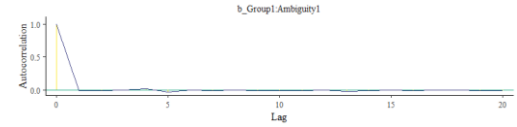
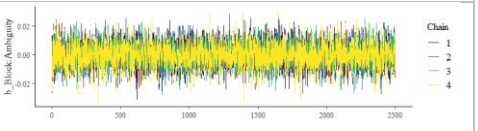
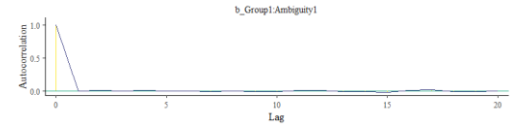
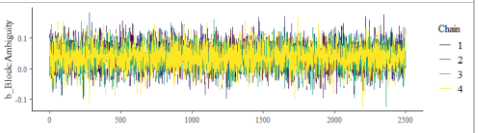
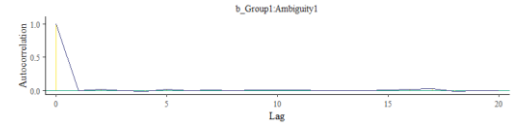
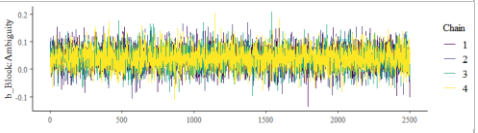
Informative		
Normal (0,1)		
Normal (0,100)		
Uniform (-10,10)		

Figure 3: Autocorrelation and trace plots for Question 2, Block and Ambiguity interaction

Now we can move on to the parameter estimates of our models.

3.4 RESULTS

For Question 1 (whether there is a significant interaction between Ambiguity condition and Group in the third block), there was a marginal effect in the original study (*p=.05)* analyzed by linear mixed effect model. However, our models with different priors all have 95% credible intervals that contain zero, which suggest no interaction between Ambiguity and Group. For Question 2 (whether there is a significant interaction between Ambiguity condition and Block for the early group), the original study didn't find a significant interaction (*p = .19),* and our results confirm the null effect. The estimate for the parameter of interest, standard error, and its 95% credible intervals are summarized as followed:

Table 4: Posterior summary using different priors for Question 1

Question1: Ambiguity*Group	Estimate (mean)	Est.Error	Lower limit of 95% Credible Interval	Upper limit of 95% Credible Interval
Informative	0.0081	0.0095	-0.0105	0.0269
Normal (0,1)	-0.0138	0.0301	-0.0724	0.0448
Normal (0,100)	-0.0141	0.0305	-0.0747	0.0448
Uniform (-10,10)	-0.0146	0.0304	-0.0752	0.0449

Table 5: Posterior summary using different priors for Question 2

Question 2 Ambiguity*Block	Estimate (mean)	Est.Error	Lower limit of 95% Credible Interval	Upper limit of 95% Credible Interval
Informative	-0.0001	0.0083	-0.0161	0.0161
Normal (0,1)	0.0388	0.0379	-0.0366	0.1124
Normal (0,100)	0.0398	0.0376	-0.0352	0.1141
Uniform (-10,10)	0.0397	0.0376	-0.0338	0.1135

Even though the substantial results are the same from models with different priors regardless of how informative they are, there are some interesting patterns found in the posterior distribution of the parameter of interest. It is obvious that from Table 4 and Table 5, the estimates of models using results from previous research as priors have the opposite sign from the other three; that is, in Question 1, only the model that used the informative prior estimates a positive effect for the interaction, while the other three models have a negative estimate of the interaction. For Question2, the model that used informative priors yields a negative estimate, whereas the other three have positive interactions.

Figure 4 further reflects that all three non/weak informative priors generate similar posterior distributions that almost completely overlap with one another, while the informative prior shows an opposite trend and deviates from the other three. Even though the substantial results are similar for these models, the model using the informative prior gives the same positive/negative trend as its prior, which favors Fine et al. (2013)'s theory, while the other three models' posterior distributions (starting from a neutral point zero) do not have such a trend.

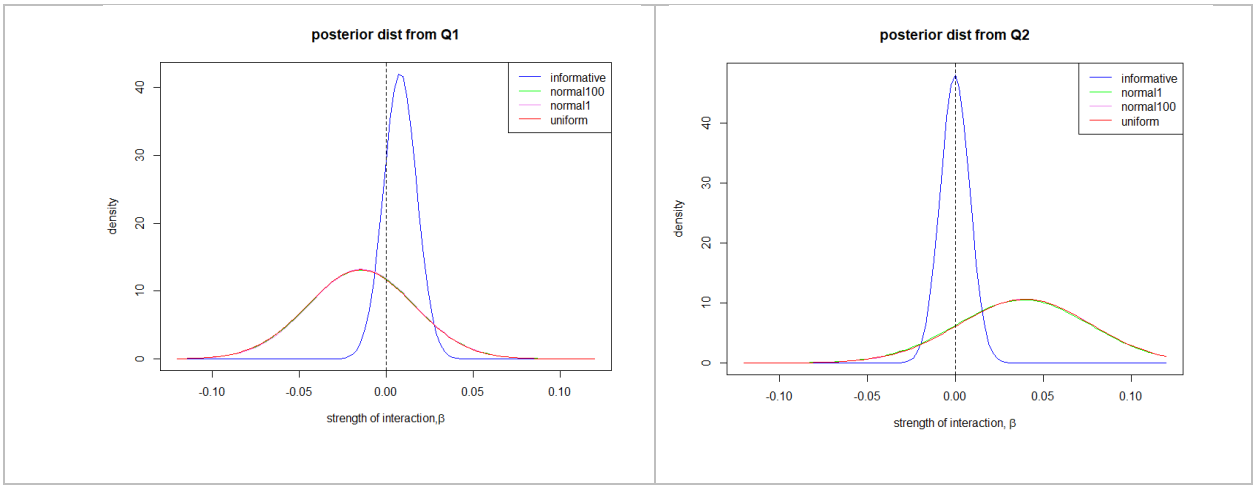


Figure 4: Posterior Distribution for Question 1 and Question 2

CHAPTER 4: MODEL COMPARISON

4.1 POSTERIOR PREDICTIVE CHECKING

To check the predictive accuracy of the fitted model, we ran a posterior predictive check with 100 replications for each model. The results and graphs are summarized in the following Figures.

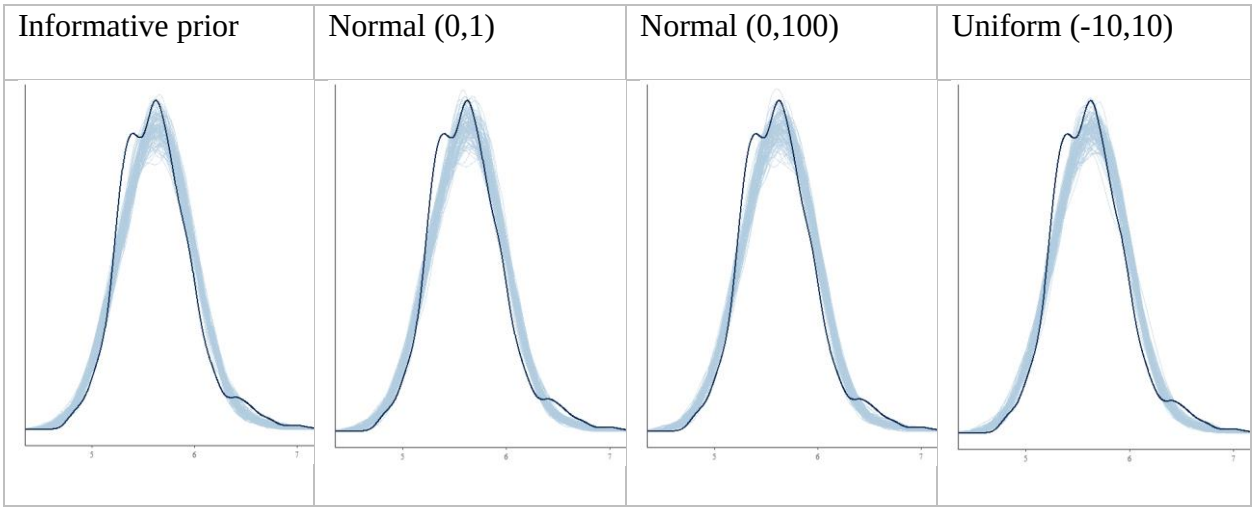


Figure 5: Posterior predictive checking for Question 1

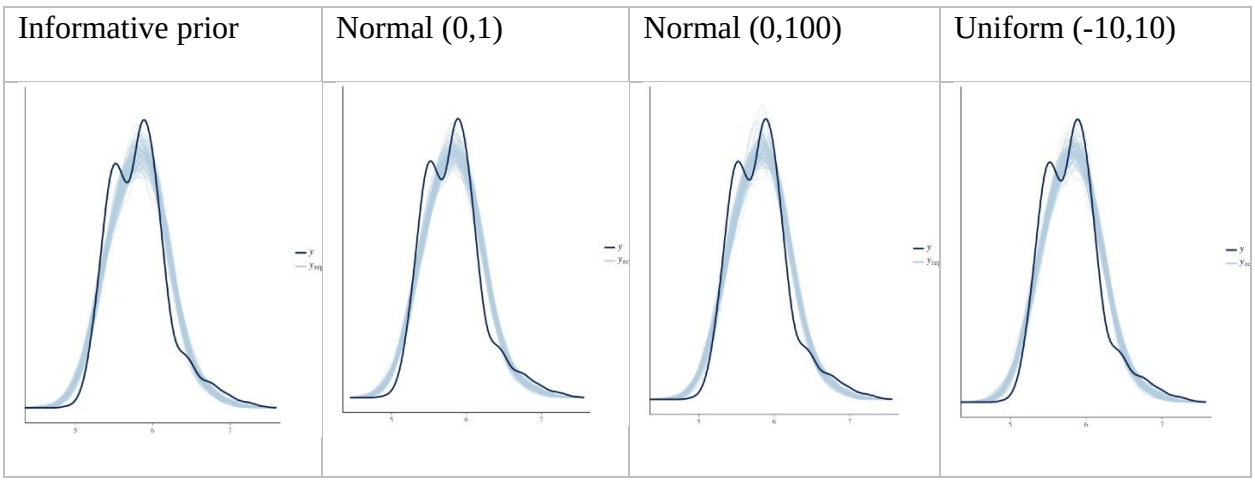


Figure 6: Posterior predictive checking for Question 2

From the graphs, we notice that all models successfully explain most of the variance in the data but fail to capture the bimodal distribution of y (the log reading time for critical regions).

The data follows a bimodal distribution due to the fact that within the same block the disambiguating verb in the ambiguous condition is coded as the critical region for both ambiguous and unambiguous stimuli, thus leading to two different mean reading times. However, since the mean reading time for these two types of sentences are very close (difference <0.1), our models fail to capture this difference if we assume it to be a unimodal gaussian. If we use a mixture of gaussian models to fit the data, the posterior distribution seems to be able to capture more of this bimodal distribution (see Figure 7: Q2, uniform prior, mixture of gaussians).

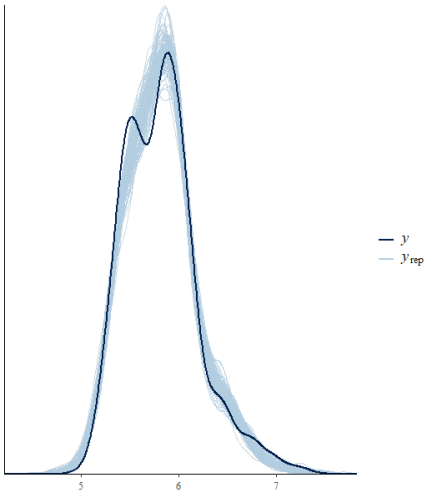


Figure 7: Posterior predictive checking for mixture models

The mixture model takes much more time and capacity to converge, and we need a sparser model for the sake of interpretation. Also, the results should be essentially the same if only the priors differ across models. Therefore, we will still use a unimodal model for the following analysis.

4.2 SENSITIVITY ANALYSIS

Another potential issue is whether or not the results from different models are stable. We conducted a sensitive analysis to check whether the results would substantially change across

models with different kinds of priors and to see if the most influential prior would lead to more unstable results if the hyperparameter were adjusted.

Table 6 shows the Potential scale reduction factor (PSRF) value between models with different priors is computed and summarized, where both rows and columns correspond to four priors we are interested in. As introduced earlier, PSRF is computed from the Gelman and Rubin convergence diagnostic for two chains. If the two chains are from the same model, values beyond 1.0±0.5 would suggest nonconvergence. If we use it for two chains from different models, then values beyond 1.0±0.5 would suggest that the two chains converge to a substantially different value. As shown in the Table 6, the upper diagonal cells show PSRF between models using different priors in Question1 and the lower diagonal cells show PSRF between models using different priors in Question2. The PSRFs between informative prior and other three are larger than 1.5, while the PSRFs compared within the three weak/non informative priors are 1's.

Table 6: PSRF between different models (point estimate and upper credible interval)

Q1 **Q2**	**Informative**	**Normal (0,1)**	**Normal (0,100)**	**Uniform (-10,10)**
Informative	Point estimate: 1 Upper C.I. :1	Point estimate: 1.58 Upper C.I. :4.52	Point estimate: 1.58 Upper C.I. :4.48	Point estimate: 1.6 Upper C.I. :4.62
Normal (0,1)	Point estimate: 2 Upper C.I. :7.88	Point estimate: 1 Upper C.I. :1	Point estimate: 1 Upper C.I. :1	Point estimate: 1 Upper C.I. :1
Normal (0,100)	Point estimate: 2.05 Upper C.I. :8.15	Point estimate: 1 Upper C.I. :1	Point estimate: 1 Upper C.I. :1	Point estimate: 1 Upper C.I. :1
Uniform(-10,10)	Point estimate: 2.05 Upper C.I. :8.14	Point estimate: 1 Upper C.I. :1	Point estimate: 1 Upper C.I. :1	Point estimate: 1 Upper C.I. :1

Then, we adjust the hyperparameter for all priors to see how the posterior distribution changes, i.e., we adjusted up 0.5 and adjusted down 0.5 for the mean based on previous prior settings in terms of the parameter of interest. For the purposes of this analysis, only Question 1's models have been analyzed.

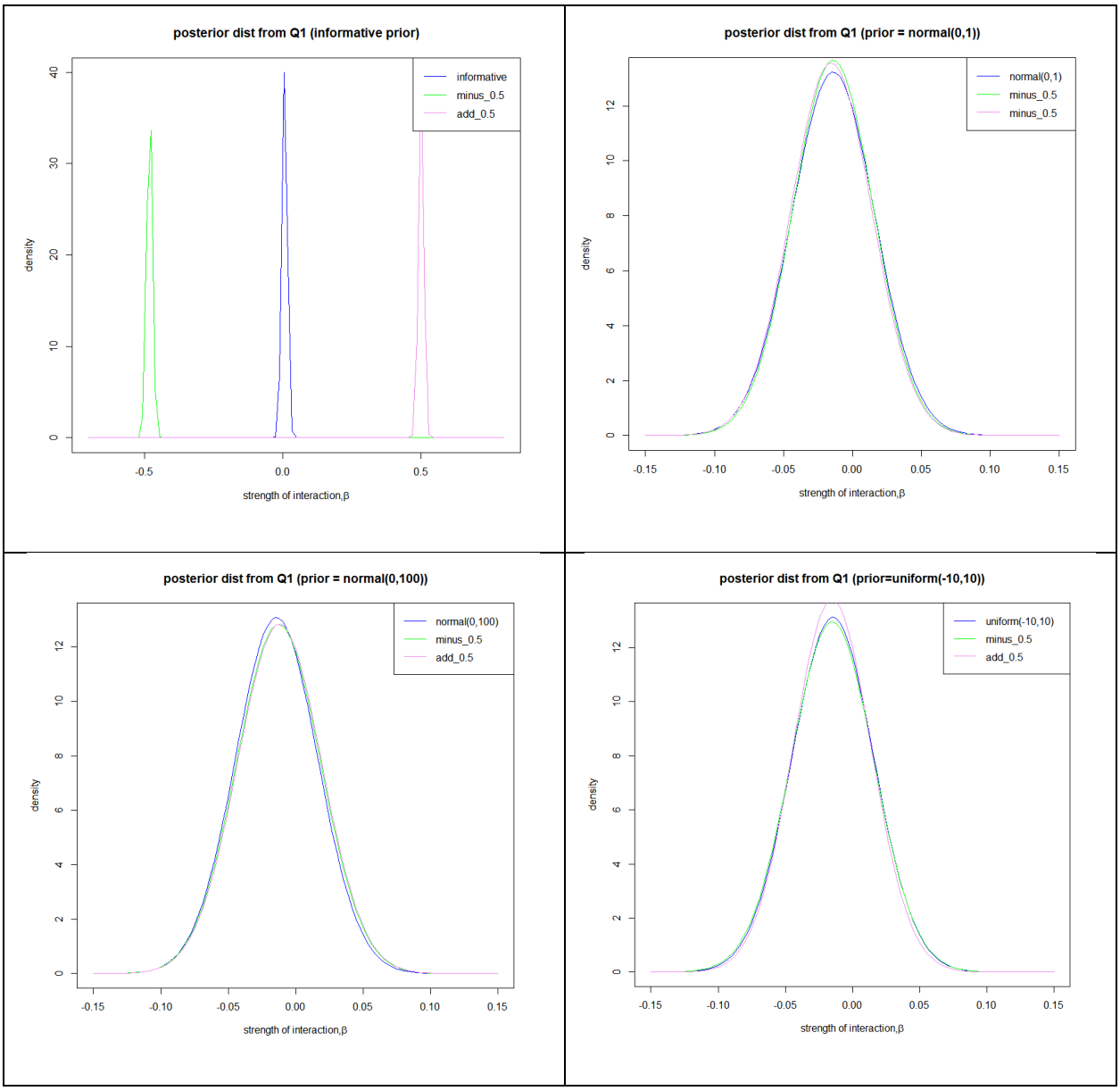


Figure 8: Posterior distribution of four models using slightly altered priors, where the posterior distribution using the original prior is depicted by the blue line, posterior distribution using the prior 0.5 lower than the original prior is depicted by the green line, and posterior distribution using the prior estimate 0.5 higher than the original prior is depicted by the purple line. Q1 corresponds to Question 1.

Apparently, only the informative prior undergoes a substantial change in terms of posterior distribution. The new estimate after a slightly (only 0.5) adjustment for informative prior would lead to such change, while it is not reflected in the posterior distribution of other models using less informative priors.

The PSRF and the percent of relative deviation are computed to further confirm the graphs (Figure 9). Given that PSRFs are much larger than 1.5 for the estimate using new priors and the relative deviation is as high as 4900%, the location of the mean hyperparameter for the prior has a large impact on the posterior.

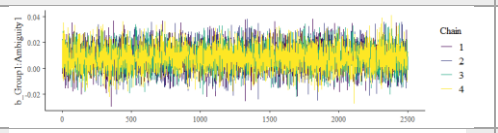
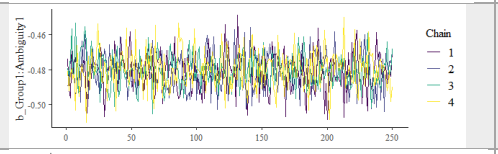
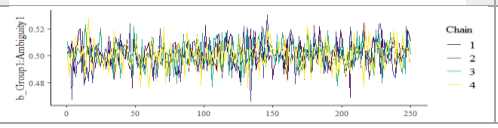
	estimate (SD)	**Trace Plot**	**PSRF**	**Size of effect (relative deviation)**
Informative prior	0.01 (0.01)			
adjust down 0.5	-0.48 (0.01)		61	4900%
adjust up 0.5	0.50 (0.01)		63.5	4900%

Figure 9: Summary of sensitive analysis for informative prior. Adjustment of priors did not lead to convergence issue (as suggested by Trace Plots). However, PSRF and relative deviation of size of the effect reflect substantial changes from the estimate using the informative prior by adding/subtracting 0.5 from the original prior settings.

4.3 MODEL COMPARISON USING INFORMATION CRITERION

Both LOO and WAIC are computed for all models and summarized in Table 7 and Table 8. Though both indices favor the Uniform prior for Question1 and the Normal (0,1) for Question 2, they do not really distinguish the models from each other given a rather large estimated standard error.

Table 7: LOO for Question 1 and Question 2

Q1	Informative	Normal (0,1)	Normal (0,100)	Uniform (-10,10)
estimate	163.3	163.4	163.3	**163.1**
SE	48.6	48.7	48.7	48.7
Q2	Informative	Normal (0,1)	Normal (0,100)	Uniform (-10,10)
estimate	-43.5	**-43.8**	-43.8	-43.8
SE	56	56	56	56.1

Table 8: WAIC for Question 1 and Question 2

Q1	Informative	Normal (0,1)	Normal (0,100)	Uniform (-10,10)
estimate	163.7	163.8	163.8	**163.5**
SE	48.6	48.7	48.7	48.7
Q2	Informative	Normal (0,1)	Normal (0,100)	Uniform (-10,10)
estimate	-43.2	**-43.5**	-43.4	-43.5
SE	56	56	56	56

Although the information criteria are not useful for model comparison in this case, they are very useful tools and should always be considered when conducting model comparisons.

CHAPTER 5: DISCUSSION AND CONCLUSION

5.1 DISCUSSION

Because priors are considered to be a vehicle for expressing psychological theory, it is suggested that using informative priors helps to advance knowledge and progress science cumulatively (Vanpaemel, 2010). However, what prior should the replication study use in Bayesian analysis if we already know we were unlikely to replicate the result (suggested by linear mixed effect modeling or other frequentist approaches)?

To answer this question, we did a case study which tried to replicate Fine et al. (2013). The replication study implemented the same fixed and random effects structure as in the original but failed to find a significant result using linear mixed effects models and subsequently used a follow up Bayesian analysis to investigate the evidence ratio between the null hypothesis and the alternative hypothesis. We tested the effect of four different priors for the posterior distribution. The first prior is the informative prior specified by the original study. The second prior is a generic weakly informative prior, Normal (0, 1). The third prior is a weakly informative prior, i.e., Normal (0, 100) where 0 is the mean and 100 is the standard deviation. The last prior is a flat prior that is traditionally considered as non-informative Uniform (-10, 10).

The results show that the directions of the target effects are different: in Question 1, only models using informative priors yielded a positive estimate, while in Question 2, only models using an informative prior yielded a negative estimate. However, the substantive conclusion given by models using different priors are all the same; that is, the credible intervals all include zero. In general, different priors need not yield the same substantive conclusions as they did in our case. It is possible for different priors to lead to different conclusions, especially if effects are stronger or more precise.

While the posterior predictive checking did not show an obvious discrepancy between simulations of fitted models and the data, the sensitive analysis suggests that the informative prior leads to a different estimate of parameter compared to other weak/non-informative models. Additionally, while the weak/non-informative priors are not much affected by slight fluctuations of hyperparameters, an informative prior is very sensitive to such changes such that the whole distribution moves towards the alternative hypothesis. Furthermore, potential scale reduction factor (PSRF) and relative deviation of effect sizes show a particularly strong deviation from the initial estimate, which confirms the earlier comparisons. Maybe this is because the standard error of the informative is much smaller than other three estimates. Finally, the information criteria (LOO and WAIC) show a trivial preference for uniform and generic weak informative prior over the informative prior.

With the rapid growth of the literature in the field, advancement in psychology is becoming more and more reliant on formal quantitative models. In the most utopian case, the parameters are not just random, vague numbers, but rather reasonable and scientific estimates of population quantities that are in accordance with psychological theories, expectations, assumptions, and intuitions. In such cases, theories can express, verify themselves in a prior distribution that contributes to a more tenable posterior estimate. However, when replicating some exploratory studies whose under-powered theories are still under scrutiny, informative priors should be used with caution as they have a substantial effect on the conclusion. It might reduce the effect when it is really there or boost an effect when it is actually not. In such circumstances, we suggest testing the model with different sets of informative/noninformative priors and choosing the one that yields the most consistent estimate after sensitive analysis, aligning best with the data in terms of posterior predictive checking, and favored most by the

information criteria. In this specific case, we choose from Normal (0,1), Normal (0,100) or Uniform (-10,10) to do the Bayesian analysis rather than the informative prior, given the three weak/non-informative priors are not substantially different from each other, and the informative prior is influential and deviates from the posterior distribution. Most importantly, because in this case we already found a nonsignificant result from the linear mixed effect model, using the informative prior, especially in Question 1, would bias the results of the replication study. Therefore, a less informative prior would help to give a more unbiased and independent conclusion in this replication study.

5.2 CONCLUSION

To conclude, when selecting priors for replication studies, we should consider several issues:

1. Did our model converge successfully judging by trace plot and autocorrelation?
2. Does the result replicate the previous study?
3. Does it make a difference to use informative, generic weakly informative, weakly informative, or noninformative priors?
4. Does posterior predictive checking mimic the data?
5. Is the posterior estimate stable if we change the entire prior distribution or slightly adjust the hyperparameter of the prior distribution (according to posterior density graphs, PRSF, and reduced deviation of size of effect)?
6. Do information criteria prefer informative/less informative models?

Rather than suggesting a single best prior selection for all replication studies, this paper aims to provide a set of guidelines for reference during replication studies when selecting priors for Bayesian analysis. If the Bayesian model did not converge in the first place, then talking

about prior selections would be meaningless because there might be a fundamental issue with model specification. If we know that the result is highly likely to replicate the previous research and using informative/less informative priors would not yield substantively different estimates, then using an informative prior specified by previous research would help to advance and solidify the theory. However, if replication is not likely (perhaps informed by a frequentist approach), and the difference between using informative/less informative priors is non-negligible, we should make a systematic comparison of models using different prior settings. Conducting a sensitive analysis will further assess how fluctuations in informative priors' hyperparameters might influence the final parameter estimate. If even the slightest fluctuation in hyperparameters leads to a great deal of instability in substantive conclusions, this might be a hint for misidentification of parameters in the model. In addition, posterior predictive checking graphs and information criteria will give a quick and intuitive suggestion of the most accurate model.

Researchers using Bayesian analyses should always be aware of the importance of prior selection and be cautious about the mismatch between data and theory. Afterall, the purpose of doing replication studies is to test, validate, and generalize theory under different circumstances and with different participants, and this verification would be spurious if the theory already biased the estimate embedded in prior distributions.

REFERENCES

Ando, T., & Tsay, R. (2010). Predictive likelihood for Bayesian model selection and averaging. *International Journal of Forecasting*, *26*(4), 744-763.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian analysis*, *1*(3), 385-402.

Bever, T. G. (1970). The cognitive basis for linguistic structures. *Cognition and the development of language*, *279*(362), 1-61.

Bijak, J., & Wiśniowski, A. (2010). Bayesian forecasting of immigration to selected European countries by using expert knowledge. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *173*(4), 775-796.

Brown, L. D. (2008). In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. *The Annals of Applied Statistics*, 113-152.

Bürkner, P. C. (2016). brms: Bayesian regression models using Stan. *R package version 0.10. 0*.

Candel, M. J., & Winkens, B. (2003). Performance of empirical Bayes estimators of level-2 random parameters in multilevel analysis: A Monte Carlo study for longitudinal designs. *Journal of Educational and Behavioral Statistics*, *28*(2), 169-194.

Darnieder, W. F. (2011). *Bayesian methods for data-dependent priors* (Doctoral dissertation, The Ohio State University).

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological methods*, *18*(2), 186.

Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(3), 327-351.

Depaoli, S., van de Schoot, R., van Loey, N., & Sijbrandij, M. (2015). Using Bayesian statistics for modeling PTSD through Latent Growth Mixture Modeling: implementation and discussion. *European Journal of Psychotraumatology*, *6*(1), 27516.

Depaoli, S., & Van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological methods*, *22*(2), 240.

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid monte carlo. *Physics letters B*, *195*(2), 216-222.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PloS one*, *8*(10), e77661.

Fransman, W., Van Tongeren, M., Cherrie, J. W., Tischer, M., Schneider, T., Schinkel, J., ... & Tielemans, E. (2011). Advanced Reach Tool (ART): development of the mechanistic model. *Annals of occupational hygiene*, *55*(9), 957-979.

Geisser, S., & Eddy, W. F. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, *74*(365), 153-160.

Gelman, A., Bois, F., & Jiang, J. (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association*, *91*(436), 1400-1412.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gelman, A., Meng, X. L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 733-760.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, *2*(4), 1360-1383.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8-38.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, 382-401.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593-1623.

Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof of the pudding: an illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological methods*, *5*(3), 315.

Ibrahim, J. G., Chen, M. H., & Sinha, D. (2014). B ayesian Survival Analysis. *Wiley StatsRef: Statistics Reference Online*.

Kaplan, D., & Depaoli, S. (2013). Bayesian statistical methods. *Oxford handbook of quantitative methods*, 407-437.

Johnson, V. E. (2013). Uniformly most powerful Bayesian tests. *Annals of statistics*, *41*(4), 1716.

Kim, S. Y., Suh, Y., Kim, J. S., Albanese, M. A., & Langer, M. M. (2013). Single and multiple ability estimation in the SEM framework: A noninformative Bayesian estimation approach. *Multivariate behavioral research*, *48*(4), 563-591.

Martin, T. G., Burgman, M. A., Fidler, F., Kuhnert, P. M., Low-Choy, S. A. M. A. N. T. H. A., McBride, M., & Mengersen, K. (2012). Eliciting expert knowledge in conservation science. *Conservation Biology*, *26*(1), 29-38.

Morris, D. E., Oakley, J. E., & Crowe, J. A. (2014). A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software*, *52*, 1-4.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of markov chain monte carlo*, *2*(11), 2.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Ostarek, M., Joosen, D., Ishag, A., De Nijs, M., & Huettig, F. (2019). Are visual processes causally involved in "perceptual simulation" effects in the sentence-picture verification task?. Cognition, 182, 84-94.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rabe-Hesketh, S., & Skrondal, A. (2012). Multilevel and Longitudinal Modeling Using Stata , Vol II.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, *83*(2), 251-266.

Rietbergen, C., Klugkist, I., Janssen, K. J., Moons, K. G., & Hoijtink, H. J. (2011). Incorporation of historical data

in the analysis of randomized therapeutic trials. *Contemporary clinical trials*, *32*(6), 848-855.

Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, *59*(4), 731-792.

Rommers, J., Meyer, A. S., & Huettig, F. (2013). Object shape and orientation do not routinely influence performance during language processing. *Psychological science*, *24*(11), 2218-2225.

Seaman III, J. W., Seaman Jr, J. W., & Stamey, J. D. (2012). Hidden dangers of specifying noninformative priors. *The American Statistician*, *66*(2), 77-84.

Stan Development Team (2018). "RStan: the R interface to Stan." R package version 2.17.3, http://mc-stan.org/.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*(1), 5-20.

van de Schoot, R., & Depaoli, S. (2014). Bayesian analyses: Where to start and what to report. *The European Health Psychologist*, *16*(2), 75-84.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*(6), 491-498.

Vehtari, A., Gelman, A., & Gabry, J. (2016). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. *R package version*, *1*(0).

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413-1432.

Vehtari, A., & Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural computation*, *14*(10), 2439-2468.

Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*, 142-228.

Zwaan, R. A., & Pecher, D. (2012). Revisiting mental simulation in language comprehension: Six replication attempts. *PloS one*, *7*(12), e51382.

Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes of objects. *Psychological science*, *13*(2), 168-171.

Wasserman, L. (2000). Asymptotic inference for mixture models by using data-dependent priors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(1), 159-180.

APPENDIX A: IRB LETTER

ILLINOIS
Office for the Protection of Research Subjects

IORG0000014 • FWA #00008584

Notice of Approval: New Submission

March 6, 2018

Principal Investigator	Kiel Christianson
CC	John Dempsey; Liu Qiawen
Protocol Title	Bayesian Prior Adjustment in Parsing (aka BPAP)
Protocol Number	18614
Funding Source	Unfunded
Review Type	Exempt
Review Category	Exempt 2
Status	Active/Data Analysis Only
Risk Determination	No more than minimal risk
Approval Date	03/06/2018

This letter authorizes the use of human subjects in the above protocol. The University of Illinois at Urbana-Champaign Institutional Review Board (IRB) has reviewed and approved the research study as described.

Exempt protocols are approved for a five year period from their original approval date, after which they will be closed and archived. Researchers may contact our office if the study will continue past five years.

The Principal Investigator of this study is reponsible for:

- Conducting research in a manner consistent with the requirements of the University and federal regulations found at 45 CFR 46.
- Requesting approval from the IRB prior to implementing modifications.
- Notifying OPRS of any problems involving human subjects, including unanticipated events, participant complaints, or protocol deviations.
- Notifying OPRS of the completion of the study.

Office for the Protection of Research Subjects
University of Illinois at Urbana-Champaign
(217) 333-2670
irb@illinois.edu

APPENDIX B: SUPPLEMENTARY FILES

Figure 10 and Figure 11 include the probability density and trace plots for each parameter in Question 1 and Question 2, showing convergence for all the parameters.

Figure 10: Probability density and Trace plots for all variables in different models, Question 1.

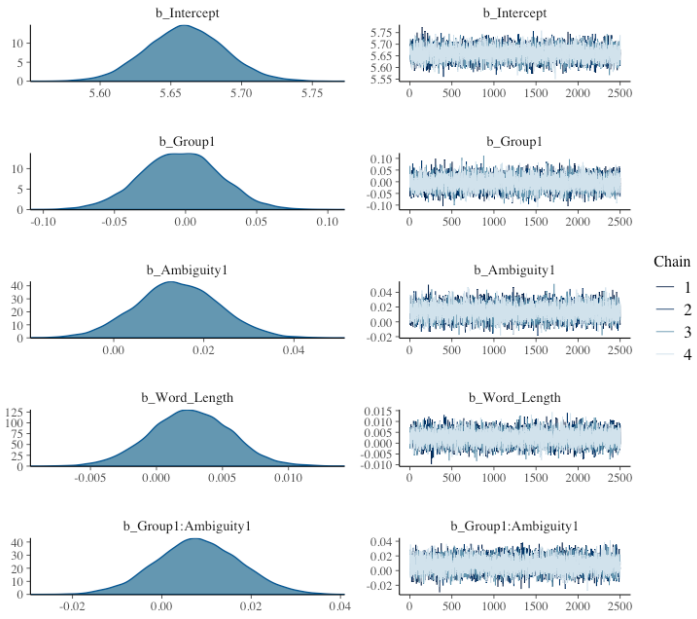
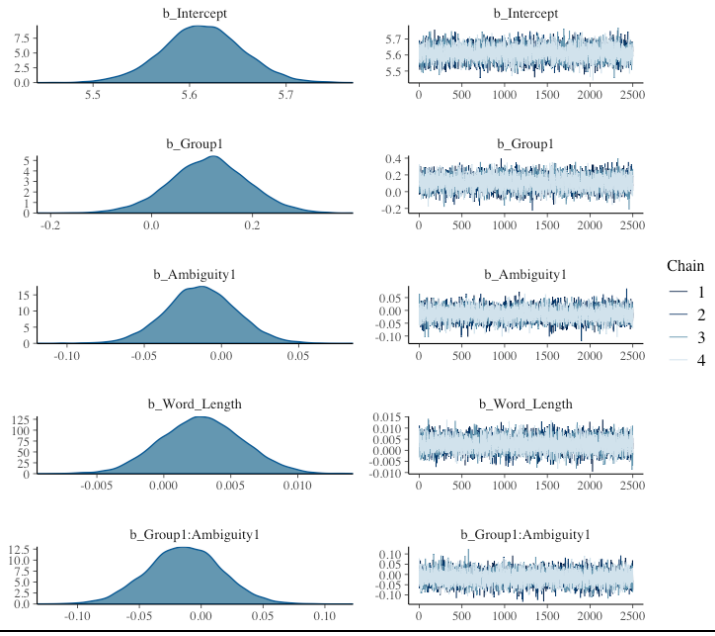
Informative	
Normal (0,1)	

Figure 10 (cont.)

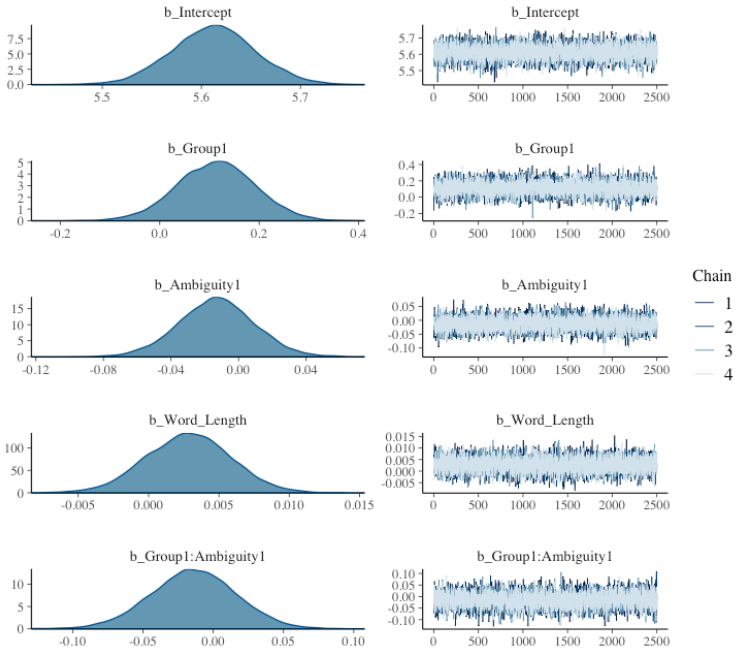
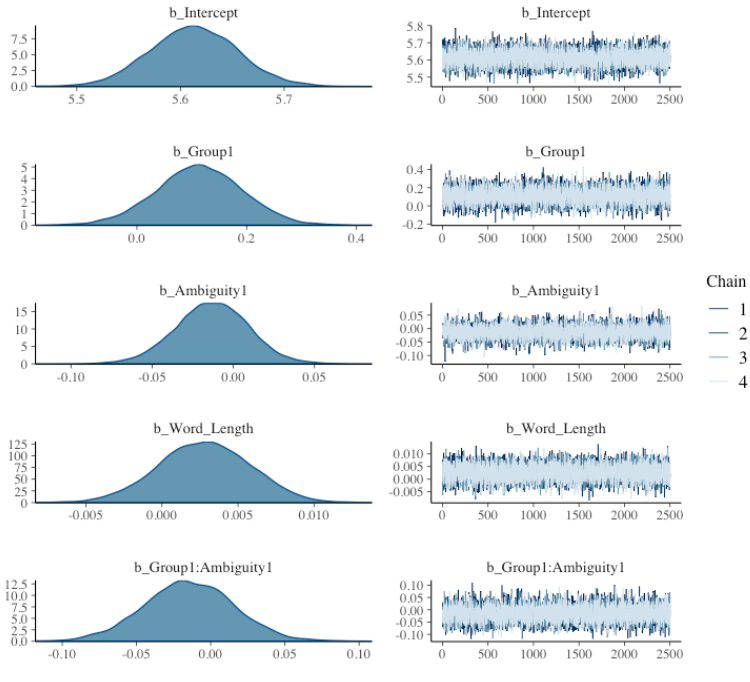
Normal (0,100)	
Uniform (-10,10)	

Figure 11: Probability density and Trace plots for all variables in different models, Question 2.

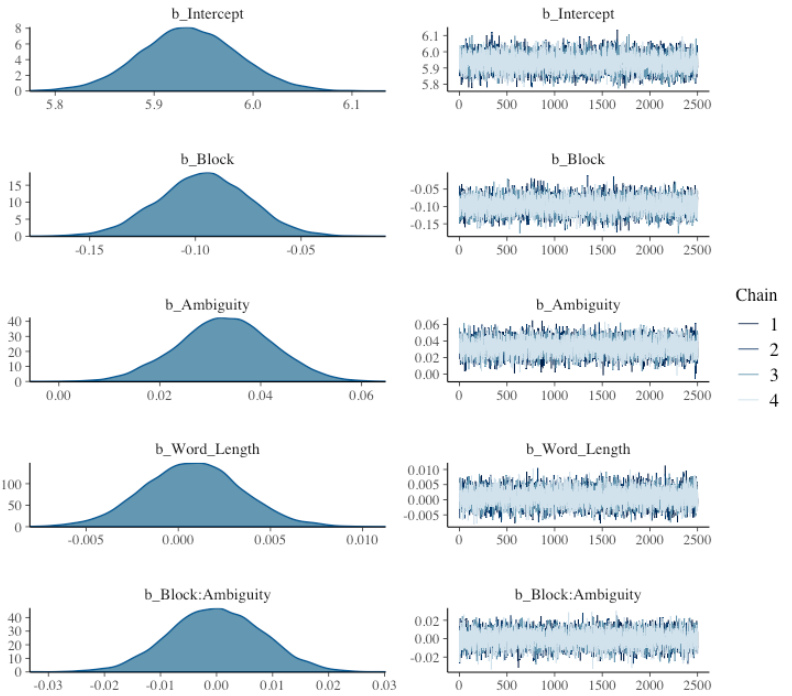
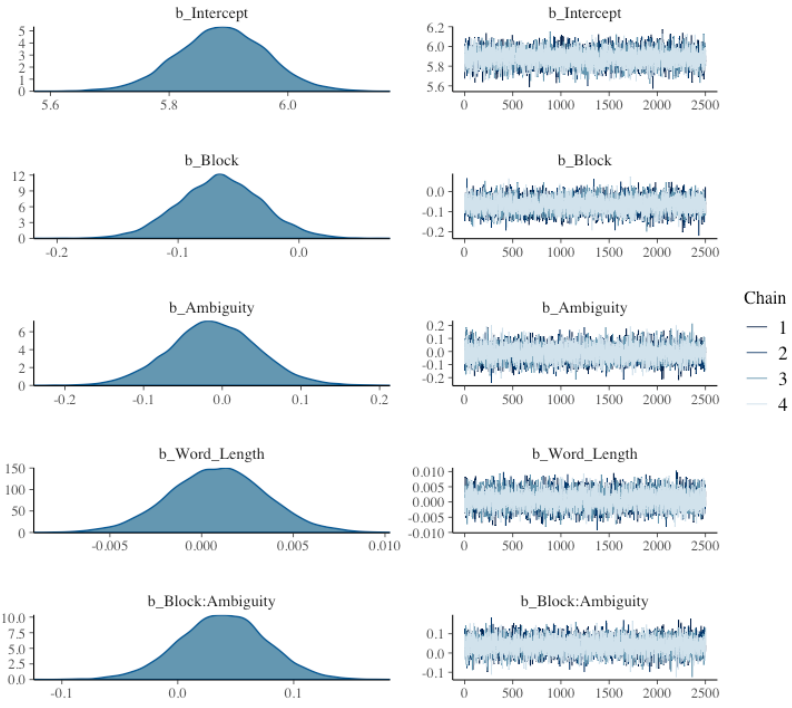
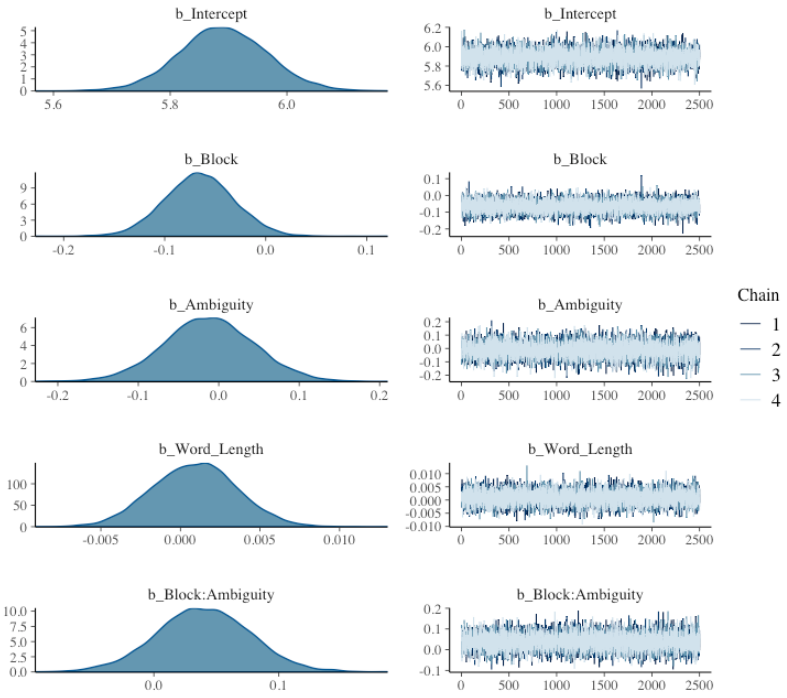
Informative	
Normal (0,1)	

Figure 11 (cont.)

Normal (0,100)	
Uniform (-10,10)